Report3

Load the R packages

```
library(knitr)
library(readr)
library(tidyverse)
library(dplyr)
library(tidyr)
library(stringr)
library(purrr)
library(httr)
library(DBI)
```

Preparations: connect to the database and get the individual tables

```
politicians <- dbConnect(RSQLite::SQLite(), "zh_politicians.db")

addresses    <- DBI::dbGetQuery(politicians, "SELECT * FROM ADDRESSES;")
affiliations <- DBI::dbGetQuery(politicians, "SELECT * FROM AFFILIATIONS;")
mandates     <- DBI::dbGetQuery(politicians, "SELECT * FROM MANDATES;")
persons      <- DBI::dbGetQuery(politicians, "SELECT * FROM persons;")
```

Part 1

Do some data cleaning: when the current politician is still active (i.e. MANDATE_END_YEAR is equal to 0), we set the year to 0. Remove politicians where either the MANDATE_START_YEAR oder MANDATE_END_YEAR is equal to zero.

```
mandates_cleaned <- mandates %>%
  filter((MANDATE_START_YEAR > 0) & (MANDATE_END_YEAR > 0))
```

Pull the start and end year to compute all the years the politicians were active:

```
start_dates <- mandates_cleaned %>% pull(MANDATE_START_YEAR)
end_dates   <- mandates_cleaned %>% pull(MANDATE_END_YEAR)
range_dates <- map2(start_dates, end_dates, seq)
```

We use the *list-colum* function to get all the mandates per year:

```
mandates_short_and_cleaned <- mandates_cleaned %>%
  mutate(FULL_YEARS = range_dates) %>%
  unnest(FULL_YEARS) %>%
  select(FULL_YEARS, ASSEMBLY)
```

Then we need to summarize all the mandates per year and assembly type:

```
mandates_short_and_cleaned_summarized <- mandates_short_and_cleaned %>%
  group_by(FULL_YEARS, ASSEMBLY) %>%
  summarize(TOTAL_MANDATES = n())
```

Now we do the plot:

```
ggplot(mandates_short_and_cleaned_summarized, aes(x=FULL_YEARS, y=TOTAL_MANDATES, color=ASSEMBLY)) +
  geom_line()
```

Part 2

Do the initial data cleaning for the PERSONS table of the database and rename the ID column for the merge in the next step:

```
persons_cleaned <- persons %>%
  filter((GENDER == 'w') | (GENDER == 'm')) %>%
  rename(PERSON_ID = ID)
```

Now merge the relevant data of the two tables:

```
mandates_persons_cleaned <- mandates_cleaned %>%
  full_join(persons_cleaned)
```

Re-do the steps from Part 1 for this merged table (there are certainly more elegant ways to copy the result from Part 1 instead of re-doing the calculations - but I decided on this approach to make sure the results are correct due to the individual data cleaning of the tables):

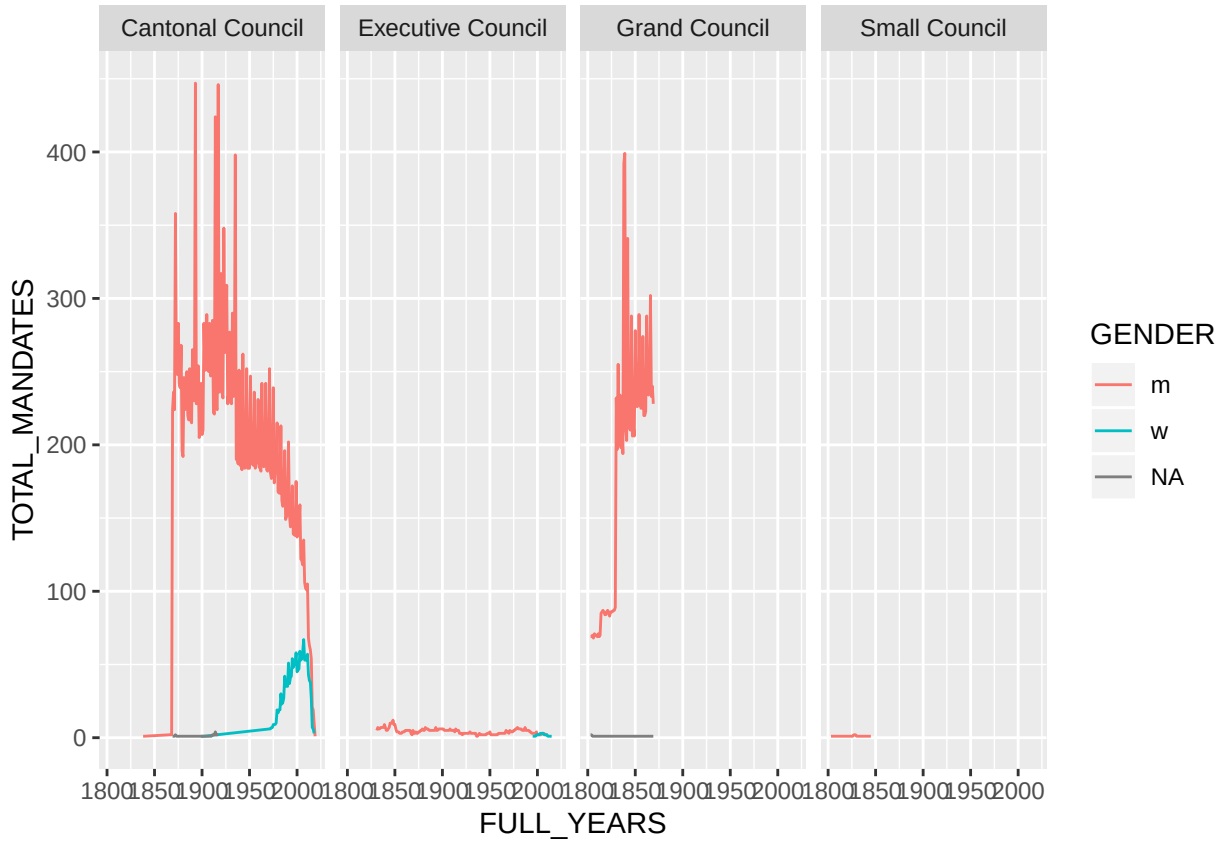
```
# Cleaning negative mandates
mandates_persons_cleaned <- mandates_persons_cleaned %>%
  mutate(DIFFERENCE = (MANDATE_END_YEAR-MANDATE_START_YEAR)) %>%
  filter(DIFFERENCE>=0)

# Unnesting and selecting the relevant columns
rm(start_dates, end_dates, range_dates)
start_dates <- mandates_persons_cleaned %>% pull(MANDATE_START_YEAR)
end_dates   <- mandates_persons_cleaned %>% pull(MANDATE_END_YEAR)
range_dates <- map2(start_dates, end_dates, seq)
mandates_persons_short_and_cleaned <- mandates_persons_cleaned %>%
  mutate(FULL_YEARS = range_dates) %>%
  unnest(FULL_YEARS) %>%
  select(FULL_YEARS, ASSEMBLY, GENDER)
```

Then we need to summarize all the mandates per year and assembly type:

```
mandates_persons_short_and_cleaned_summarized <- mandates_persons_short_and_cleaned %>%
  group_by(FULL_YEARS, ASSEMBLY, GENDER) %>%
  summarize(TOTAL_MANDATES = n())
```

Now we do the plot:

```
ggplot(mandates_persons_short_and_cleaned_summarized,aes(x=FULL_YEARS, y=TOTAL_MANDATES, color=GENDER))
  geom_line() +
  facet_grid(cols=vars(ASSEMBLY))
```



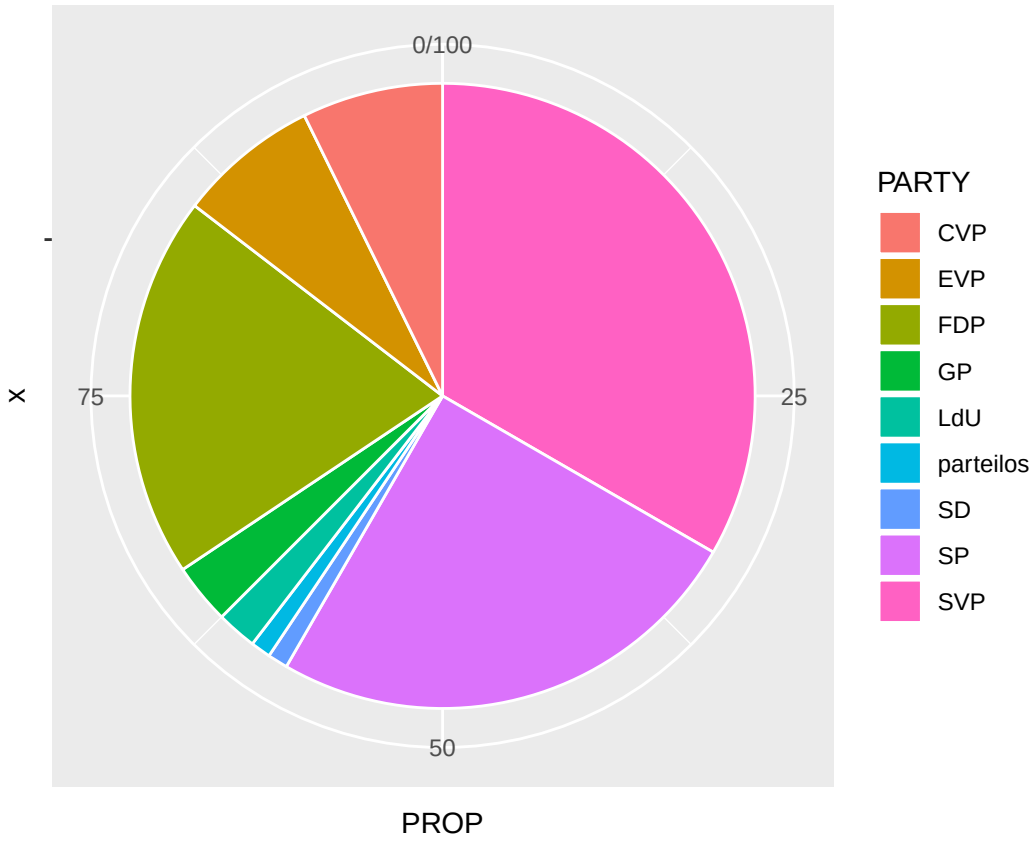
Part 3

Start with the usual data cleaning and data unnesting:

```
affiliations_cleaned <- affiliations %>%
  mutate(PERIOD = (AFFILIATION_END_YEAR - AFFILIATION_START_YEAR)) %>%
  filter(PERIOD > 0)

rm(start_dates, end_dates, range_dates)
start_dates <- affiliations_cleaned %>% pull(AFFILIATION_START_YEAR)
end_dates   <- affiliations_cleaned %>% pull(AFFILIATION_END_YEAR)
range_dates <- map2(start_dates, end_dates, seq)

affiliations_2000 <- affiliations_cleaned %>%
  mutate(FULL_YEARS = range_dates) %>%
  unnest(FULL_YEARS) %>%
  filter(FULL_YEARS==2000) %>%
  select(ID, PARTY) %>%
  group_by(PARTY) %>%
  summarize(SEATS = n()) %>%
  mutate(PROP = 100*SEATS / sum(SEATS))
```

Now we do the plot:

```
ggplot(affiliations_2000, aes(x = "", y = PROP, fill = PARTY)) +
  geom_bar(width = 1, stat = "identity", color = "white") +
  coord_polar("y", start = 0)
```



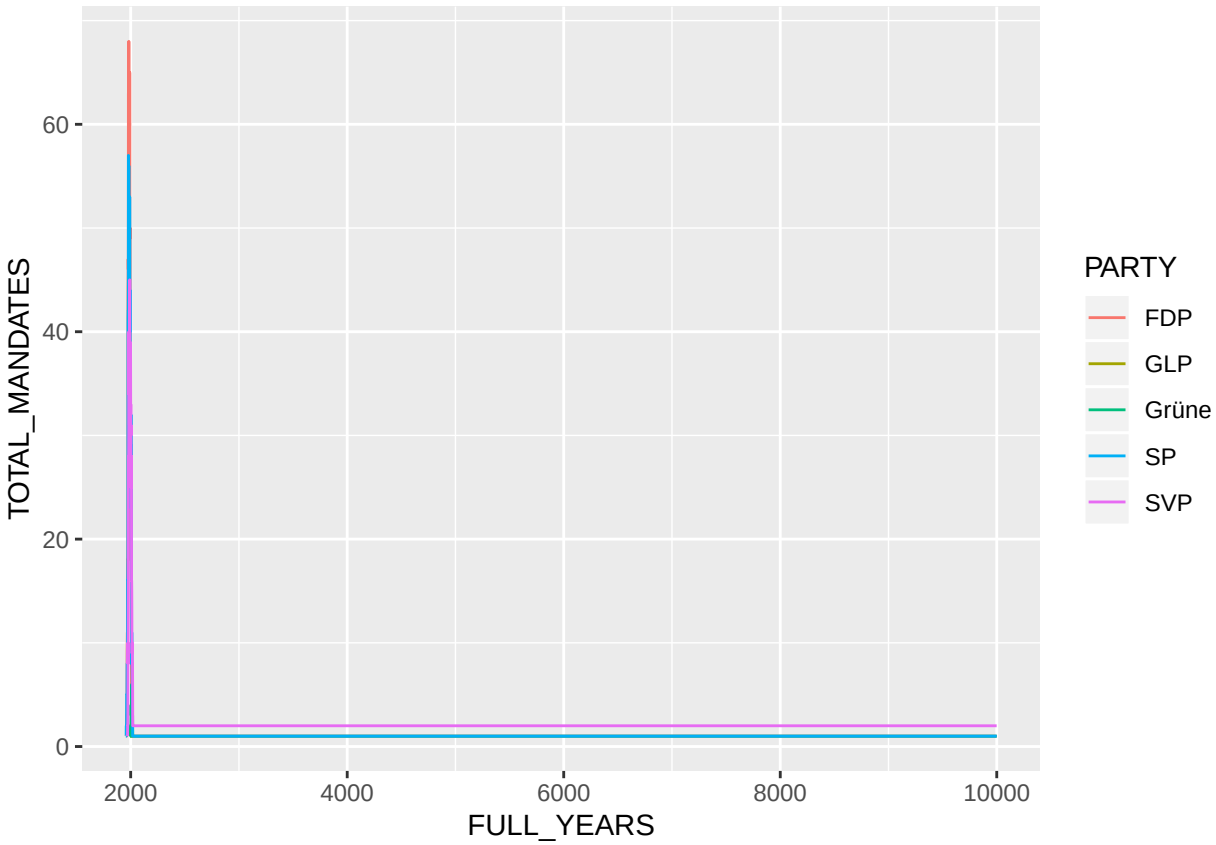
Part 4

Start again with data manipulation tasks:

```
affiliations_cleaned_and_summarized <- affiliations_cleaned %>%
  mutate(FULL_YEARS = range_dates) %>%
  unnest(FULL_YEARS) %>%
  group_by(FULL_YEARS, PARTY) %>%
  summarize(TOTAL_MANDATES = n())
```

Now we do the plot but select only a couple of parties:

```
  affiliations_cleaned_and_summarized %>%
  filter((PARTY=="SVP") | (PARTY=="SP") | (PARTY=="Grüne") | (PARTY=="GLP") | (PARTY=="FDP")) %>%
  ggplot(aes(x=FULL_YEARS, y=TOTAL_MANDATES, color=PARTY)) +
  geom_line()
```


TOTAL_MANDATES
60
40
20
0
2000
4000
6000
8000
10000
FULL_YEARS
PARTY
FDP
GLP
Grüne
SP
SVP